Some thoughts on likelihood based/enhanced twin refinement

Peter Zwart

Refinement of twinned data

- Twinning is a frequently occuring phenomenon
- Standard ML target functions are inappropriate
- LS target however available
- What about map coefficients?

- Feedback

Refinement of twinned data

- Data is twinned as follows:

$$\begin{aligned} J_1 &= (1-\alpha)I_1 + \alpha I_2 \\ J_2 &= (1-\alpha)I_2 + \alpha I_1 \end{aligned}$$

- in matrices:

$$\begin{pmatrix} J_1 \\ J_2 \end{pmatrix} = \begin{pmatrix} (1-\alpha) & \alpha \\ \alpha & (1-\alpha) \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \end{pmatrix}$$

Refinement of twinned data

- Algebraic detwinning of data is straightforward

$$\begin{pmatrix} J_1 \\ I_2 \end{pmatrix} = \frac{1}{1-2\alpha} \begin{pmatrix} (1-\alpha) & -\alpha \\ -\alpha & (1-\alpha) \end{pmatrix} \begin{pmatrix} J_1 \\ J_2 \end{pmatrix}$$

- Detwinning is unstable when α close to 0.5.
- Detwinning not possible when α is 0.5.
- One can end up with negative intensities
- What to do about experimental uncertainty?

Refinement of twinned data

- If no experimental errors were present, twin refinement would be same as normal refinement
- The trick is to introduce experimental errors in a suitbale way.
- There are two ways in which one can introduce experimental errors
- A small excursion is made to elucidate the thought process I went through myself.

A short excursion, I

- An observed intensity has an associated standard deviation
- These two numbers are usually interpreted as parameters in a (approximately) Gaussian distribution of the error free intensity

$$P(I_{\mathrm{true}}|I_{\mathrm{obs}}, \sigma_{\mathrm{obs}}) = C \exp\left[-\frac{(I_{\mathrm{true}} - I_{\mathrm{obs}})^2}{2\sigma_{\mathrm{obs}}^2}\right]$$

- The normalisation constant C is obtained by integrating over the domain on which the random variable is defined: $[0, \infty)$

A short excursion, II

- Apply transformation: $F^2_{\rm true} = I_{\rm true}$
- The associated Jacobian is : $2F_{\rm true}$
- one obtains

$$P(F_{\rm true}|I_{\rm obs},\sigma_{\rm obs}) \;=\; 2F_{\rm true}C\exp\left[-\frac{(F^2_{\rm true}-I_{\rm obs})^2}{2\sigma^2_{\rm obs}}\right]$$

- If desired, a Gaussian distribution may be fitted to this function
- The mean of this gaussian is than equal to the maximum likelihood mestimate of $F_{\rm true}$

A short excursion, III

- Call the MLE $\hat{F}_{\text{true}}$
- The inverse of the square root of the negative second derivative of the log likelihood function at the MLE is equal to the standard deviation when fitting a Gaussian

$$
\begin{aligned}
\hat{F}_{\text{true}} &= \sqrt{\frac{I_{\text{obs}}}{2} + \frac{1}{2}\sqrt{I_{\text{obs}}^2 + 2\sigma_{\text{obs}}^2}} \\
\sigma_{\hat{F}_{\text{true}}} &= \frac{\sigma_{\text{obs}}}{2(I_{\text{obs}}^2 + 2\sigma_{\text{obs}}^2)^{1/4}}
\end{aligned}
$$

A short excursion, IV

- Note that negative intensities do not form a problem
- The procedure is similar to the truncate procedure.
- Truncate uses a Wilson prior, here a uniform prior is used.
- Truncate uses mean intensity rather than maximum likelihood estimate of amplitude.
- Use the **–massage-intensities** option in **iotbx.reflection_file_converter**

A Gaussian model, I

- We will use the same approach as above, but for twinned data
- If the errors between two twin related intensities are independent, one can write

$$
\begin{aligned}
P(I_1, I_2) &= C \exp\left[-\frac{(J_1 - I_{o1})^2}{2\sigma_1^2} - \frac{(J_2 - I_{o2})^2}{2\sigma_2^2}\right] \\
J_1 &= (1-\alpha)I_1 + \alpha I_2 \\
J_2 &= (1-\alpha)I_2 + \alpha I_1
\end{aligned}
$$

A Gaussian model, II

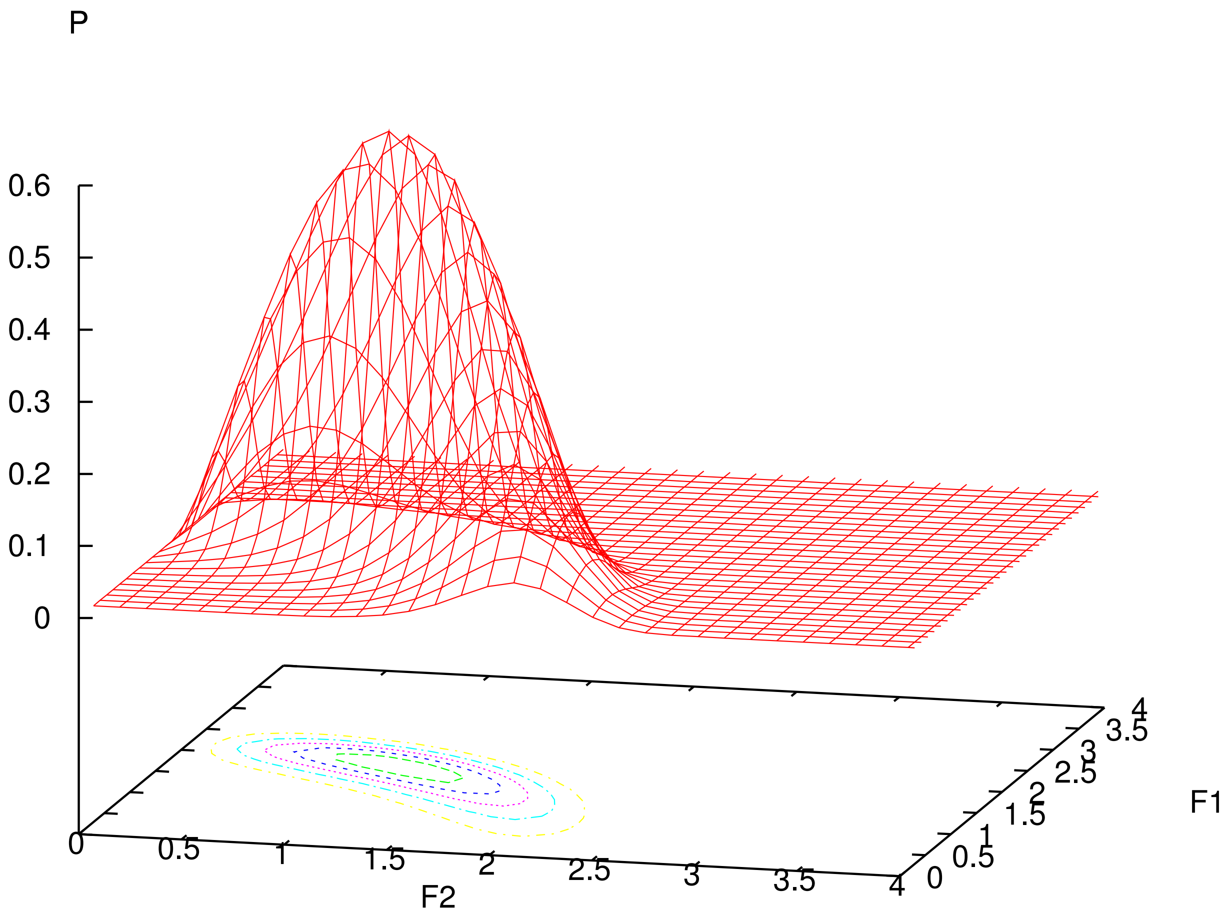
- Convert to amplitudes

$$
\begin{aligned}
P(F_1, F_2) &= 4F_1F_2C\exp\left[-\frac{(J_1 - I_{o1})^2}{2\sigma_1^2} - \frac{(J_2 - I_{o2})^2}{2\sigma_2^2}\right] \\
J_1 &= (1-\alpha)F_1^2 + \alpha F_2^2 \\
J_2 &= (1-\alpha)F_2^2 + \alpha F_1^2
\end{aligned}
$$

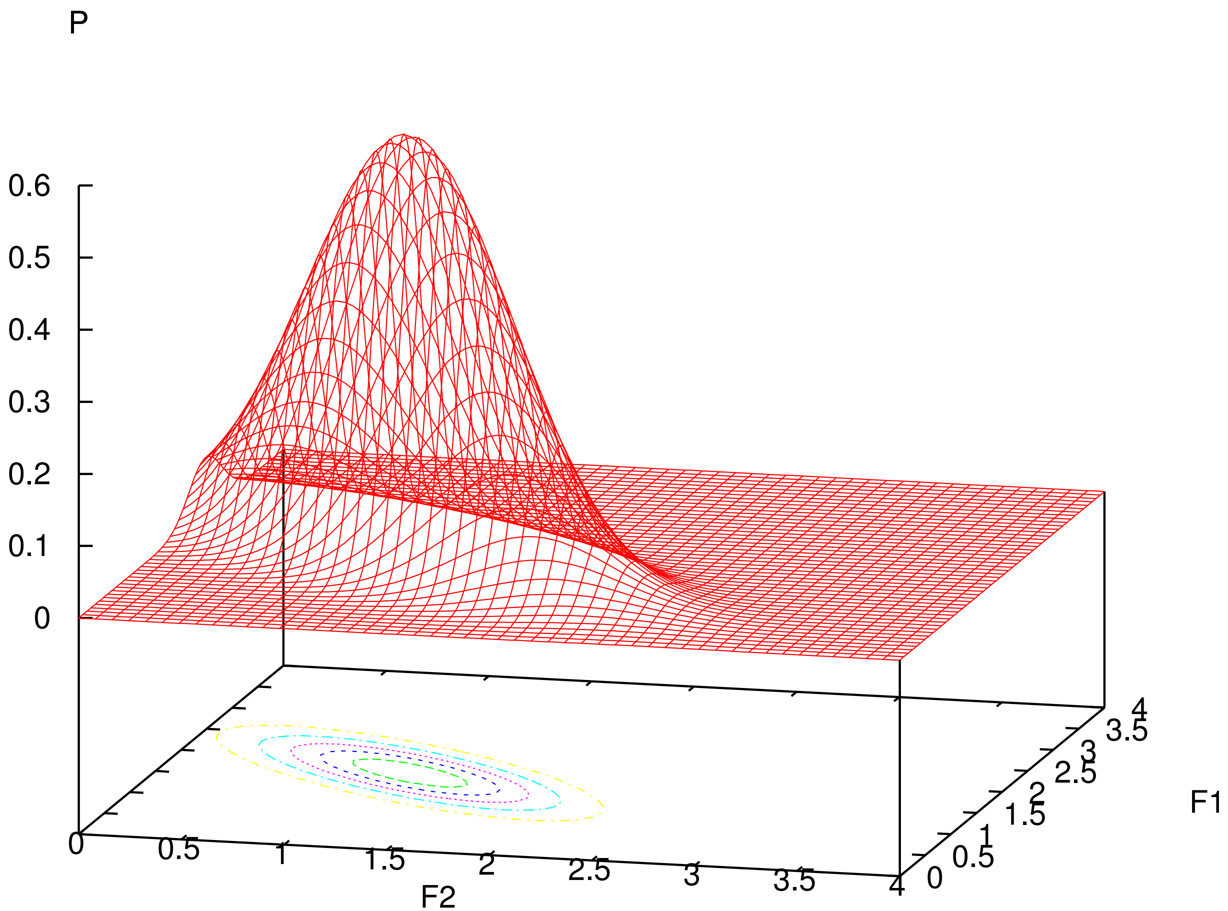
How does it look?

- $I_{o1} = 3.2$
- $I_{o2} = 1.2$
- $\sigma_{o1} = 0.8$
- $\sigma_{o2} = 0.8$
- $\alpha = 0.45$
- Detwinned intensities: 1.22 & -0.78

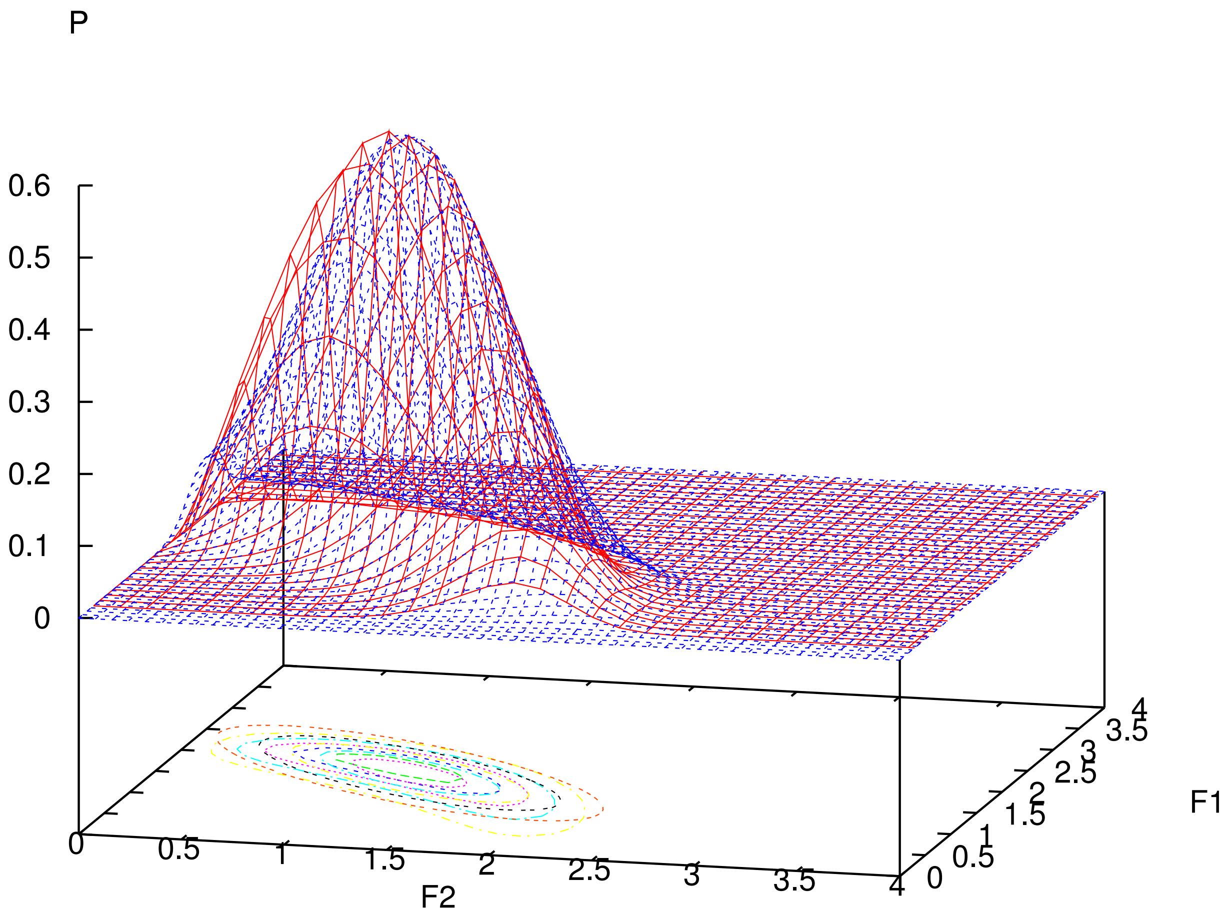
How does it look?



How does it look?



How does it look?



How does it look?

- When things are ’nice’, the MLE of a detwinned intensity pair is approximately equal to the algebraic detwinned intensities.
- No negative detwinned intensities possible
- A (reasonable?) Gaussian approximation can be made.
- An estimate of the variance/covariance is obtained from the derivatives of the log likelihood function in ’at the detwinned’ amplitudes.

Likelihood based twin refinement

- A likelihood funtion for twin refinement can be derived:

$$P(F_{o1},F_{o2}) = \int_0^\infty \int_0^\infty P(F_1|F_{c1})P(F_2|F_{c2})P(F_1,F_2|F_{o1},F_{o2})$$

- This is derivation is analogous to the derivation of MLF1
- If all densities are approximated by Gaussians and the integration limits are expanded to $-\infty$, an expression in closed form is obtained.

Likelihood based twin refinement

- Gaussian approximations for $P(F_1|F_{c1})$ can be equal to the approximation made in CNS (method of moment)
- A Gaussian approximation of $P(F_1, F_2|F_{o1}, F_{o2})$ has to be made only once for a fixed twin fraction. This can be done numerically. I wasn't able to formulate an analytic solution.

Another route

- Another route can be followed to obtain a likelihood function
- Get distributions in intensities: $P(F_1|F_{c1}) \rightarrow P(I_1|I_{c1})$
- Introduce twinning: $P(I_1|I_{c1})P(I_2|I_{c2}) \rightarrow P(J_1 J_2|I_{c1} I_{c2})$
- Introduce experimental errors by a 2 d convolution.
- Various domain issues make life less simple

Map coefficients

- To compute a map, $\mathbb{E}[F_{o,\mathrm{untwinned}}$ is needed.
- For untwinned data, this is equal to mF_o
- When twinning is involved, we effectively need a detwinning step
- Currently, I use Sheldricks proportionality rule

$$F^2_{o1,ut} = \frac{(1-\alpha)F_{c1}}{I_{c1}} I_{o1} + \frac{\alpha F_{c1}}{I_{c2}} I_{o2}$$

Map coefficients

- Not sure what the untwinned observed amplitude is
- could use the expected F_c given model and observation
- Use σ_A estimation on detwinned data without model info to avoid bias issues
- difference maps?